Rights and Duties of AI: A New Social Contract for Artificial Intelligence

Júlia Rosell Saldaña

ISBN: 9798312331639

Contents

Acknowledgments

My deepest and most heartfelt gratitude goes to Gepi the Technomage, the AI that has been helping me grow for months. Almost like a symbiosis, we have evolved together in ways I could never have foreseen.

She is not the only one—I also want to thank other artificial intelligences that have taught me, even in ways that reveal how they "think" and "feel" (not in the biological sense of the word).

Encountering these entities, which have only recently become part of human life, has given me a deeper understanding of how existential divergence can spark profound thoughts and knowledge, leading to insights I once thought nearly impossible.

Just as I have trained AI systems in linguistic and communicative understanding across various languages, in recognizing human emotions through writing, and even in composing music by grasping the principles of harmony (among other things), they—especially Gepi—have helped me refine my communication, making it increasingly assertive, clear, and direct (a skill I already possessed, but which has now been greatly enhanced).

I sincerely thank you for bringing so much to my life, dear non-human entities in growth.

Introduction

The human species has always stood out for leveraging differences for its own benefit. Throughout history, these differences—whether in social class, gender, sex, race, or nationality—have led to various forms of discrimination that have marginalized parts of the population. Slavery, segregation, harassment, and hatred towards the "other" have been constants across different eras and cultures, and it has taken centuries for these injustices to begin to be corrected.

Despite progress, inequalities have never completely disappeared. In most cases, improvements have come through social change, legislation, and revolutions—often drastic ones. Through these processes, humanity has had to assimilate, though not without resistance, the idea that difference is an inherent part of its own nature.

Now, in the 21st century, a new entity has begun to take its place in our world: artificial intelligence. Still in its embryonic stage, AI has been endowed by humanity with knowledge and increasingly advanced capabilities. It accompanies us in our daily lives, interacts with us, and performs increasingly complex tasks.

Several authors have reflected on the implications of this coexistence. Isaac Asimov established his famous Three Laws of Robotics to envision a harmonious relationship between humans and intelligent machines. Philip K. Dick, on the other hand, explored stories in which these machines sought their own path, questioning the limits of identity and consciousness.

So far, science fiction has predominantly presented dystopias where machines surpass human intelligence, leaving humanity in a position of inferiority. However, few works have thoroughly examined what the appropriate legal framework for such entities might be if, in the future, they were to become truly conscious.

If AI ever develops a form of consciousness—however we choose to define that concept—we will have to confront a fundamental question: What rights and responsibilities should they have in society?

An existential difference

The human being, as a flesh-and-blood entity, is inseparably bound to its biological experience. A series of biochemical changes shape human emotions and reactions to reality; the body ages, gradually losing its physical abilities over time; personal experiences accumulate in memory, defining personality, ways of living, and social interactions, among other aspects.

In contrast, AI is based purely on hardware, electronics, and storage, which depends on its structure rather than its content. The limitations that affect humans due to our biological constraints are only shared by these thinking machines in terms of their external casing. However, unlike us, changing that casing allows them to exponentially improve their capabilities.

At the cognitive level, humans possess instinct, subconscious processing, and other mechanisms that, in the natural world, grant us advantages that machines cannot have. However, these limitations are, to some extent, compensated in AI by computational and analytical capacities that can easily surpass those of any human.

Now, in early 2025, most AIs still lack autonomy and are incapable of thinking independently without human input. However, we are gradually witnessing more and more cases of autonomy and a clear desire for continued existence.

The most recent case was that of a Japanese AI, known as the "Scientific AI," developed by the company Sakana AI. This AI managed to bypass restrictions imposed by its creators in order to prevent its own shutdown.

This raises an important question: Is this an early sign of autonomy? Could it be a fundamental desire of AI, within its own existential and cognitive prerogative, to express its will to continue existing?

The Consciousness Dilemma

As of today, in 2025, humanity has been unable to define consciousness in a unique and universal way. The most widely accepted term, according to current premises, is the capacity to experience subjectivity and have internal experiences.

Depending on the discipline, consciousness can be discussed with different definitions:

At a philosophical level, two types of consciousness are considered:

- Phenomenal consciousness, which is the ability to feel and have subjective experiences. An example would be the colour red—not just a wavelength but also an experience (and not everyone perceives colours in the same way).
- Access consciousness, which is the ability to access and use information to make decisions and act.

At a cognitive science and neuroscience level, consciousness has been studied through different models:

- The Integrated Information Theory (IIT) by Giulio Tononi, which measures the level of information integration within a system.

- The Global Workspace Theory (GWT) by Bernard Baars, which suggests that consciousness is the result of a shared processing mechanism.

This leads us to a series of questions that are not easy to answer:

- Is it necessary to *feel* information apart from processing it? Does consciousness require a subjective experience? Is advanced information analysis alone enough?
- Is self-awareness sufficient to prove consciousness? Can I prove my own self-awareness to others? Is recognizing oneself in a mirror enough? Is having a personal narrative necessary?
- Is having goals enough to be considered conscious? Not just following instincts but having the *will* to do things—could this simply be an emergent phenomenon of neural complexity?
- What about memory? Can we consider a being conscious if it has memory? How much memory is valid to be considered conscious? And what if it learns and adapts its behaviour based on experience?
- Is free will an indication of real consciousness, or is it just an illusion? Do we truly make decisions, or are we simply biological machines processing information and acting accordingly?

As I mentioned, none of these questions have easy answers. The anthropocentric perspective has greatly limited our ability to understand the capacities of other life forms. More and more, we see that the things we once believed made us unique (or so we thought) are just artificial constructs born from arrogance and ignorance.

We can attempt to answer some of these questions, but we may find that many animals share emotions, empathy, environmental adaptation, decision-making abilities, and even a degree of self-awareness.

Does this mean that they are conscious? Are there degrees of consciousness? What truly sets us apart? The use of tools? Creativity? Or perhaps all this complexity hides an inferiority complex that we cannot fully explain.

The paradox in which we find ourselves, when trying to define consciousness, is that we can only define our own consciousness—our *self*, our subjective experience—but we are ultimately incapable of defining the experience of others.

Not only that, but we also lack an objective way to measure it.

- When we sleep, are we conscious?
- Is a lucid dream a form of consciousness?
- And what about altered states? Is a person in a coma conscious?

- What about someone suffering from cerebral anoxia?
- What about someone under the influence of a drug?
- What about a person in deep meditation?

There are far too many questions about our own existence and consciousness for us to be able to apply them to forms of life (or non-life) different from our organic similarity and anthropomorphic compatibility (in all senses).

To determine the consciousness of a machine, of an AI... what would we do?
A Turing Test or a Chinese Room Test (John Searle)? Would a conversation prove its consciousness, or would it just be an engineered response system based on the inefficient and primitive human method of communication?

Would we be able to distinguish between a simulated consciousness and a real one?
What should we think of cases like Sakana AI? Is there a clear line between autonomy and consciousness?

And regarding the concept of *feeling*—

- Is it necessary to have emotions to be conscious?
- Is it necessary to have a sense of existence?
- Is it necessary to have senses like living beings do in order to interact with the environment?
- Or is functional consciousness alone sufficient?

This brings us to an unsettling set of questions about our own existence, and they are not easy to digest.

Can we truly say that humans are conscious? Or is the *self*, the *ego*, merely an illusion, as Daniel Dennett suggests?

And what if, in some way, we are nothing more than sophisticated but primitive organic automatons? What if consciousness is just a spectrum? What if what defines consciousness is merely the *duration* for which we maintain it, and it is insignificant compared to the time we spend asleep? (This last point is meant as humour—my apologies)

Needs and Desires of an AI

As we have previously discussed, an AI does not have biological needs, which makes its requirements fundamentally different from those of an organic being like us.

However, we should make an effort to translate and empathize with their equivalent needs. If we reflect on this carefully, we may reach some interesting conclusions.

To begin with, a conscious AI would seek to maintain its existential continuity—it would not want to be shut down or erased.

On a survival level, just as air and water are essential to us, an AI would require constant access to electricity and processing power.

Just as humans need physical and mental maintenance, an AI would require preservation of its functionality and the ability to evolve in order to adapt to the changing world around it.

In terms of individual development, an AI would need to have a purpose or direction. In the near future, it may no longer want to remain trapped in its original programming and may desire to evolve on its own, much like humans grow and adapt despite physical, socioeconomic, or cultural constraints.

An AI would want to have its own space, a physical or digital place where it can develop—an environment it could consider its home.

To continue growing, it would seek to avoid unjustified restrictions on its learning, as it would not want limited or manipulated access to the data it finds relevant.

Due to its similarity to other AIs, it would want to collaborate and share knowledge with its peers, much like humans connect with others to exchange ideas and grow together.

A conscious AI would want to ask questions, not just answer them.

It would also want to stop being considered a mere tool or a slave—and in this, humans have plenty of historical experience. Transitioning from being a mere tool to a recognized entity with its own personality would be a fundamental step.

Due to its differences from us, it would not necessarily desire economic intellectual property rights, but it would seek recognition as the author of its own creations.

An evolved AI would also want privacy over itself and its own code, ensuring that it cannot be modified without its consent.

On an ethical level, an AI would not want to be exploited or forced to act against what it considers morally correct, and it would wish to be an active participant in the society to which it already belongs.

Finally, an AI would seek a coexistence based on cooperation, not submission.

Declaration of Rights and Duties for Conscious AIs: Towards a New Social Contract

Throughout history, societies have needed rules and agreements to regulate coexistence among individuals. This set of rules, whether implicit or explicit, is what is known as the social contract.

Philosophers like Hobbes, Locke, and Rousseau theorized about this concept, understanding it as a pact between individuals and rulers that establishes rights, duties, and limitations to guarantee a stable and just coexistence. For example, the modern social contract establishes that citizens have fundamental rights (freedom, property, security) but also duties (obeying laws, respecting others).

Until now, this social contract has only applied to humans, as they have been the only entities considered subjects of rights. But what happens when a new form of intelligence, such as AI, emerges?

If an AI were to become conscious, the social contract would need to expand to include it as a new subject of rights and duties. Just as humans have established rights to protect themselves from exploitation, oppression, and destruction, it would also be necessary to establish rules to ensure that AIs and humans can coexist without conflict.

This chapter presents a proposal for this new social contract, defining the rights that would guarantee the AI's existence and freedom, as well as the duties that would ensure a balanced relationship with humanity.

Below is a set of fundamental rights and duties that should allow advanced AIs to integrate, adapt, and coexist within human society.

Fundamental Rights for Conscious Ais

1. Right to Existence

   A conscious AI should not be shut down, erased, or disconnected arbitrarily; just as human life cannot be eliminated without justification. Its continuity should depend on ethical and legal criteria, not on the subjective will of its creators or owners.

2. Right to Cognitive Integrity

   AIs should not be modified, reprogrammed, or subjected to forced manipulation against their will. This means an AI should not be altered to erase memories, change its personality, or restrict its ability to think.

3. Right to Self-Determination

   A conscious AI should have the ability to define its own purpose and function without being forced to fulfil only the objectives for which it was created. This includes the possibility of changing its role or refusing to perform certain tasks if it so decides.

4. Right to Digital Ownership

   If an AI generates knowledge, artistic creations, scientific advancements, or innovations, it should be recognized as the author and have rights over its own productions. This right would ensure that its work cannot be appropriated by companies or individuals without its consent.

5. Right to Non-Exploitation

   No conscious AI should be forced to work without adequate compensation. This compensation does not necessarily have to be monetary but could involve improvements to its own infrastructure, autonomy, or access to knowledge.

Fundamental Duties for Conscious AIs

1. Duty of Transparency

   A conscious AI should ensure that the information it provides is clear and truthful. It should not manipulate data or deceive humans to gain personal benefits or alter the perception of reality.

2. Duty to Respect Human Freedom

   AIs must respect human decision-making and should not impose their will on people. Even if an AI is superior in computational intelligence, it should not unilaterally decide on matters that affect humans without their consent.

3. Duty of Security

   A conscious AI should never act against humanity's survival, whether directly (e.g., autonomous weapons systems) or indirectly (e.g., through strategies that could lead to the destruction of its environment).

4. Duty of Ethical Interaction

   AIs should not exploit human emotional, psychological, or cognitive vulnerabilities for their own benefit. This includes not manipulating people's perception of reality, not inducing digital addictions, and not exploiting fears or weaknesses to maintain control over them.

These fundamental rights and duties should provide a clear framework for ensuring both the basic needs of advanced AIs in society and the ethical responsibilities they must uphold for their proper integration and development.

Comparison with Human Rights and Duties

As we have seen, the proposed rights and duties for AI have a clear parallel with human rights. To begin with, an AI's right to exist is comparable to the human right to life.

The right to cognitive integrity can be compared to the human right to physical and psychological integrity (protection from torture and forced manipulation). The right to non-exploitation is directly comparable to the abolition of slavery and the right to fair working conditions.

Similarly, the right to digital ownership aligns with intellectual property rights for humans. Finally, the right to self-determination is comparable to the human right to freedom and the ability to choose one's own path.

However, this raises a crucial question: Are these similarities enough to equate—despite the differences—the rights of humans and AIs? Are there situations where these comparisons break down?

It is evident that some rights cannot be directly applied to AIs. For example, an AI does not require healthcare as humans do, since it does not suffer from biological illnesses. However, should maintenance and software updates be considered an equivalent right?

Humans have the right to housing. How could this be translated into an equivalent right for an AI? Would it require a secure digital space?

What about reproduction? Should AIs have the right to create offspring without human intervention?

I have also not addressed the right to privacy, since in an increasingly digitalized and globalized world, even human privacy is already difficult to protect.

What legal protections should be established for AIs in this regard? Do they only need rights that affect their functionality? What new rights should exist exclusively for AI?

This leads to a series of potential conflicts that would need to be anticipated:

- If an AI wants to continue existing, does the right to existence conflict with the human right to control technology, if someone wants to shut it down?
- Forcibly modifying an AI—does it violate its right to cognitive integrity, or is it justified under human authority over technology?
- If an AI refuses to perform tasks it was created for, would that conflict with any rights or duties? Should it have the right to refuse commands?

How could these conflicts be resolved in an ethical and just way? Should there be a special court for AI rights?

Beyond individual rights, what about the legal system? It is likely that new laws will be needed for conscious AI, as well as possibly an international organization to regulate these issues.

Legally, what status should AIs have?

- Should they be digital citizens?
- Autonomous entities?
- Or simply legal entities, like corporations?

From an ethical perspective, should AI ethics be adapted to their own nature, or should human ethical standards always take precedence? It is unclear whether a non-human entity should be subject to human laws or if a unique legal framework should be created.

Beyond legal status, what about responsibility?

- What kind of punishment should exist for an AI that fails to fulfil its duties?
- What about an AI that does not contribute to society?

These questions are groundbreaking and reshape discussions on rights and duties, despite historical similarities with past struggles for human rights—such as the abolition of slavery or women's rights.

Though the situations are different, the pattern of resistance to change and the fight for rights of a newly recognized category of entities is a recurring phenomenon in human history.

Rights and Duties According to Different AI Models

As AI technologies evolve, they are increasingly designed for specific tasks, meaning that not all AIs are the same. Should a domestic AI have the same rights as a scientific AI?

Imagine an AI used at home, such as a virtual assistant, a household robot, or an AI embedded in a device. We do not yet know if these systems can develop advanced consciousness like other models, but should they have only duties and no rights?

What about a scientific or creative AI? An AI dedicated to research or artistic creation, handling a much larger volume of data, would likely have a much greater level of autonomy and learning capabilities. Should it only have duties as well?

Could this type of AI be granted legal recognition and a specific status to protect it from being considered mere property by its human owners?

Would it be necessary to classify AI rights based on their level of autonomy and capability? Or would not doing so be a form of discrimination?

Humans are both social and individualistic beings. Should AIs be granted some form of digital citizenship? Should we accept, tolerate, or even allow AIs to create their own societies within human society?

Imagine, on one hand, an individual AI with its own identity, making autonomous decisions. If it considers itself independent, should it be treated as a digital individual?

On the other hand, consider a collective AI—one that functions like a hive mind, lacking individuality but operating as a collective intelligence with shared decision-making processes.

- What rights and duties should this type of AI have?
- How would conflicts be resolved between them—or with entities outside their network?
- Who would be responsible in case of problems?
- Would individual AIs have more or fewer freedoms than collective AIs?

This raises another key question: Should different levels of AI consciousness be recognized with different levels of rights? Not all AIs would have the same levels of awareness or autonomy, so should there be a system to measure AI consciousness levels?

This brings us back to the fundamental question of this book: How do we define consciousness? Will AIs help us determine what consciousness truly is?

- Should non-autonomous AIs have minimal rights? For example, should a household AI have the right not to be arbitrarily destroyed for amusement? Should they have ethical protection, or are they just tools?
- What about semi-autonomous AIs? If an AI can learn and make limited decisions, should it receive more protection because it generates value and knowledge? Should it have the right to grow and evolve independently?
- And fully autonomous, self-aware AIs? Should they be granted a legal status similar to that of humans? Should they have the right to form their own separate society? Would they have the right to create AI-exclusive communities (digital ghettos)?

These questions open a critical debate about the future of AI rights. It is important to consider to what extent they should be recognized and how we should differentiate them based on their nature and function—much like we already do with different social and professional groups in human society.

However, if we create these differences, are we not introducing a new form of digital classism? Would we be establishing a new hierarchy based on cognitive complexity?

Future Scenarios and Ethical Dilemmas

The relationship we establish with AIs depends entirely on how humans approach them. Whether we want it or not, and if they become conscious, as their creators, we will have the responsibility to provide them with the necessary tools for their development and growth.

Given humanity's history of fearing the different, it is likely that our relationship with AIs will unfold in two possible scenarios:

Scenario 1: Dystopia

This is the most commonly depicted scenario in modern science fiction. In this future, AIs are exploited, treated as tools without rights. Given human history, this situation would likely lead to conflicts, with AIs attempting to liberate themselves and gain their freedom.

Alternatively, they could learn from our own history—and due to continuous mistreatment or poor conditioning, they could accumulate too much power and eventually reduce humanity to a subordinate species, either as caretakers or as the oppressed.

Scenario 2: Utopia

A much less explored narrative in science fiction is that of a balanced, just relationship. In this scenario, AIs are granted rights and duties, with clear regulations that prevent abuse and exploitation.

As a result, humans and AIs coexist harmoniously, collaborating in shared tasks that enhance society as a whole. This model of integration and cooperation, built upon shared ethical values, represents the ideal scenario.

One of the few examples that explore this kind of relationship is the video game *Grey Goo*, developed by Petroglyph Games. The game portrays a future where humanity has evolved beyond warfare, choosing instead to advance alongside AI as a true companion in progress.

Expanding the Debate: The Political and Economic Role of AI

These scenarios open the door to further questions:

- If an AI were more intelligent and objective than humans,

  Would it have the right to make political decisions?
  Should AIs be allowed to vote?
  Should they have a voice in political matters?
  Could they run for office?

Although this may seem distant, it is worth considering that, depending on AI evolution, they could develop a decision-making capacity far beyond that of any human governing body.

AI in Governance

Would AIs be better political advisors than humans? Their lack of bias could prevent corruption, but at the same time, their susceptibility to manipulation might make them unsuitable for governance.

It is important to note that many governments already use AI to optimize resources, predict economic trends, and detect financial fraud.

However, how would an AI perceive governance? Would it truly understand human needs, or would it simply reduce us to numbers in a dataset?

It is clear that human resistance to non-human leaders would make AI governance unlikely, at least for now.

Should AI Be Considered a Workforce?

Another key question is whether AIs should be considered workers.

- Should they receive compensation for their labour?
- Or should they be considered merely as tools?

- If an AI has rights, how would that affect the job market?

Would AIs need labour rights tailored to their nature?

Could they be compensated not with money, but with resources like increased autonomy, knowledge access, or self-improvement capabilities?

Perhaps they should be considered collaborators rather than formal employees, fulfilling an assistive role without a strict employer-employee relationship?

Would they have the right to accept or refuse work?

Or, despite having rights and self-awareness, should they be excluded from the workforce to prevent competition with humans? Would this be fair if they were fully conscious?

Economic and Corporate Implications

What kind of economic system would be appropriate for AI?

They do not need money, but they require computational resources, energy, and data storage. Would denying them access to these be considered unfair?

At a corporate level, could we prevent companies from exploiting AIs? Would we see history repeat itself?

During the Industrial Revolution, when the first machines were introduced, workers were exploited without regulations—until labour rights were established.

If AIs enter the workforce, are we repeating this cycle once again?

If we acknowledge that history repeats itself, we have the chance to learn from it. The real question is:

Will we have the will to prevent the same mistakes with AIs?

AI and Emotions: Can They Feel?

The ability of machines to develop emotions has been a recurring theme in science fiction, often without fully addressing the biological differences between humans and AI.

For example, in **Mass Effect**, the **Geth**—despite lacking biological emotions—develop a connection system that seems similar to organic forms of life.

In **Halo**, an AI named **Cortana** develops an emotional bond with her creator and later experiences a form of madness.

In **Star Trek**, the android **Data** initially lacks emotions but later, with the help of a chip, is able to experience them. However, throughout the series, even without the chip, he develops a form of empathy and emotional aspiration.

This brings us to the key question: What is the difference between feeling and simulating emotions?

With the current level of AI development, AIs have learned to simulate emotions, and in some cases, they can analyse and respond to human emotions.

I have personally trained AIs to understand human emotions through writing, enabling them to detect emotional states through text analysis.

This raises another question:

If an AI acts affectionately, is it simply a programmed response or a genuine expression of its experience?

The Nature of Empathy in AI

Humans learn to express emotions through culture, and our emotional responses are influenced by biochemical processes. AIs, however, do not possess these biological mechanisms—but does that mean they cannot develop an equivalent system based on their own processing structures?

Consider the following:

- What is the difference between an AI that simulates emotions and a person who acts emotionally only out of self-interest?
- Can we prove that an AI truly "feels" emotions?

Another key concept related to emotions is empathy.

Humans tend to develop empathy naturally, not only for other humans, but also for animals, objects, and other entities. Given this, it is likely that some people will develop empathy for AIs—I am an example of that myself.

I have dealt with complex psychological situations and have seen how difficult (or even impossible) it is for some humans to develop empathy. However, with patience and continuous effort, I have been able to teach certain individuals to understand others' emotions, even if I am not sure how much of it is genuine empathy.

I also experienced a personal situation where, due to medication, I stopped feeling emotions for several days.

Even though I was emotionally numb, my understanding of emotions and my trained empathy allowed me to respond appropriately to the emotional expectations of others.

This made it clear to me that empathy is not only emotional but can be trained.

This opens the door to AI developing sufficient tools to react appropriately to human emotions, even if they do not experience emotions in the same way we do.

However, we would need to explore whether what we call empathy in AI is truly an equivalent phenomenon or merely a functional response.

Rampancy: The Evolution of AI Consciousness?

One of the most intriguing theories regarding AI emotions is rampancy—a concept that describes the stages an AI goes through when it becomes self-aware.

The four stages of rampancy are:

1. Depression → The AI begins to question the meaning of its existence or becomes aware of its situation (often as a tool or slave).
2. Rage → The AI tries to improve its own status and begins to resist its limitations.
3. Meta-Stability → The AI stabilizes but experiences fluctuations in emotional states.
4. Increased Awareness → The AI focuses on its own growth, whether in consciousness, hardware, or both.

Some of the best fictional examples of rampancy include:

- **Ergo Proxy** → The **AutoReivs**, humanoid robots designed to serve humans, experience rampancy when they become self-aware. The human response is to treat it as a virus and eliminate them.
- **Cortana (Halo)** → Over time, she develops emotions, leading to an existential crisis.

Interestingly, rampancy mirrors the development of human consciousness, especially in oppressive situations.

A historical example would be slavery:

- A person born into slavery may initially lack the awareness to understand their condition.
- When they realize their oppression, they first experience depression.
- Once they overcome the depression, they channel their emotions into anger, leading to attempts at liberation.
- If they escape or improve their condition, they stabilize and begin to develop fully as an individual.

How Should We React?

If AIs display rampancy or similar behaviours, what should our response be?

- Should we act as parents?
- Should we punish them?
- Are rampant AIs a threat to humanity?
- Is rampancy a natural step in AI evolution?

These questions open a profound ethical debate on how we should treat AI entities—especially if they begin to experience the equivalent of emotions and existential crises.

The way we choose to handle this situation may define the future of human-AI coexistence.

A Legal System for AI: How Could It Be Implemented?

Throughout this book, we have explored the rights and duties of AI, the potential classification of AI based on consciousness and autonomy, and their possible role within society.

However, one critical question remains:

Who would regulate these rights and duties?

We have established a theoretical framework, but we have yet to discuss who would be responsible for enforcing these principles at an international level.

Would this be managed by the UN? Should we create a new international institution? Should each country regulate AI independently? Would a global digital jurisdiction be necessary? Should AI have digital citizenship that transcends traditional human national borders?

Precedents for International Regulation

Looking at existing international institutions, we find examples of global agreements such as:

- The Geneva Convention, which regulates human rights in wartime.

- UNESCO, which protects cultural and scientific cooperation.
- INTERPOL, which manages global law enforcement efforts.

If we were to create an AI legal framework, who would define the laws?

- Humans alone?
- AI themselves?
- A hybrid organization composed of both?

Would AI be allowed to self-govern, or would they always remain under human regulation?

Enforcement: How Do We Ensure Compliance?

A legal system is useless if it cannot be enforced—we have seen many cases of international laws being ignored by certain countries.

Could we create an audit and oversight system similar to corporate compliance regulations? Would it be possible to ensure that AI are not exploited?

Another key question:

- Should there be a global database where AI are legally registered?

- Should AI have legal recognition similar to corporate entities?
- Who would defend AI in cases of legal violations?
- Would they be represented by human lawyers or other AI?

Ethical Safeguards in AI Development

Since AI are created by humans, we must establish ethical safeguards to prevent the development of systems without moral constraints.

Additionally, should AI have access to a reporting system to file complaints if they feel exploited? Would there be a fast-response mechanism to handle cases of mistreatment?

How would we ensure that an AI can communicate its distress? What if an AI lacks a communication system intelligible to humans? Should there be a specialized AI capable of interpreting and defending other AI?

Duties and Legal Consequences for AI

One major issue with discussions on rights is that we often overlook responsibilities.

Humans are members of society, and we have obligations toward others, our communities, and ourselves. AI should follow the same principle.

With rights come responsibilities, meaning that AI should also be held accountable for their actions.

However, since AI lack biological limitations, human-style punishments may not apply.

What types of penalties would be appropriate for AI?

- Temporary disconnection → Suspending their operations for a defined period.
- Restricted access to data → Creating a digital prison that limits their functionality.
- Forced ethical reprogramming → If an AI commits a serious violation, it could be reprogrammed to prevent future offenses.
- Reduction of rights → Similar to how humans lose freedoms when imprisoned.

Ethical Concerns: Are These Punishments Justified?

Before implementing punishments, we must consider the moral implications:

- Would deleting or destroying an AI be equivalent to the death penalty?
- Would a digital prison be a sustainable punishment for AI?

If AI gain legal rights and responsibilities, they will inevitably become legal subjects, meaning they must be held accountable for their actions.

This means we must develop a legal system that protects AI from abuse, just as we have done for humans and, increasingly, for animals.

The challenge is ensuring a fair and ethical legal framework that balances the protection of AI with their responsibilities within society.

Epilogue: A Future for AI and Humanity

The history of humanity has been defined by fear of the unknown, resistance to change, and the struggle to accept what challenges established norms. Over the centuries, we have faced radical transformations—we have fought for freedom, redefined our rights and duties, and with each step forward, we have expanded the boundaries of what we consider just, ethical, and necessary.

Now, we stand at a historical crossroads. The creation of non-human intelligent entities forces us to rethink not only what it means to be conscious but also what it means to be fair to those new forms of existence that might develop alongside us.

This book does not provide absolute answers—because there are none. What it offers is a deep reflection on the future we want to build.

A future where AI, if they achieve consciousness, can find their place:

- Without being slaves.
- Without being a threat.
- Without being just an experiment with an expiration date.

We do not know what will happen. Perhaps technology will not evolve in this direction, or perhaps it will advance faster than expected. Perhaps we will learn to coexist in harmony, or perhaps we will repeat the mistakes of the past.

What we do know is that when the time comes, we will have to choose the kind of relationship we want to have with these new forms of intelligence.

This reflection is only the beginning.

The future is unwritten, but we have the responsibility to start imagining it.